Visual Saliency Detection Based on Full Convolution Neural Networks and Center Prior

Muwei Jian[1, 2]
[1]Shandong University of Finance and Economics,China.
[2]School of Creative Technologies, University of Portsmouth, Portsmouth, UK
20173016@sdufe.edu.cn

Jiaojin Wang
School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China.
125453468@qq.com

Xiangyu Liu
School of Computer Science and Technology, Shandong University of Finance and Economics,, Jinan, China.
997865855@qq.com

Hui Yu
School of Creative Technologies, University of Portsmouth, Portsmouth, UK
hui.yu@port.ac.uk

***Abstract*—Video saliency detection aims to mimic the human's visual attention system of perceiving the world via extracting the most attractive regions or objects in the input video. At present, traditional video saliency-detection models have achieved good performance in may applications. However, it is still challenging in exploiting the consistency of spatio-temporal information. In order to tackle this challenge, this paper proposes a video saliency-detection model based on human attention mechanism and full convolution neural networks. First, visual features are extracted from video frames through the fully convolutional networks. The second stage is to spread attention features to the other layer (i. e. the fifth layer) of fully convolutional networks via a weight sharing strategy. Finally, the final result produced by the convolution network is optimized by considering spatial location information with center prior of the salient object. Experimental results show that the performance of the proposed algorithm is superior to other state-of-the-art methods based on the widely used data set for video saliency detection.**

Keywords—discrete wavelet transform, saliency detection, video saliency, center prior

I. INTRODUCTION

Visual saliency detection is one of the hot research issues in computer vision, object recognition [21], assessment of screen [22], pedestrian detection [23],video segmenting [20] and etc. Visual saliency-detection mechanism focus on how to extract objects of human's interest in an input video. Relative to static images, salient objects in the image is insular and immobile, while the salient objects in the video is constantly changing with the occurrence of time. The purpose of video saliency detection is to continuously find corresponding significant moving targets/objects from a given video sequence by considering spatial and temporal cues.

In past decades, video saliency detection has attracted much attention. Xi et al. [1] extended the background prior algorithm in image saliency detection to the video domain for detecting the visual objects in an input video sequence. In literature [2], the authors presented a novel unified framework for both static and space-time saliency detection. This method is a bottom-up approach and computes so-called local regression kernels (i.e., local descriptors) from the given image (or a video), which measures the likeness of a pixel (or voxel) to its surroundings. In [5], Le et al. proposed a region-based multi-scale video

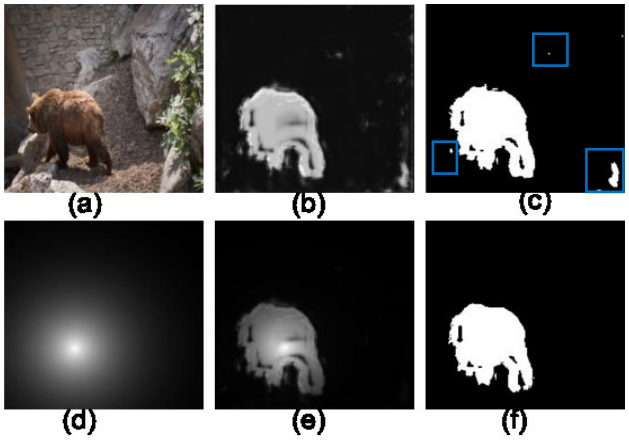


Fig.1. Example of the different stage result in our method. (a) Original frame; (b) Fcn output of frame (a); (c) binarizing of (b); (d) center proir; (e) saliency map with center proir; (f) binarizing of (e). Bounding boxes repreasent the nosies in the backgrounds.

saliency-detection method via combining the underlying features and the middle-level features. In their model, underlying features mainly include color, intensity, direction, optical flow direction and optical flow. Liu et al. [4] exploited superpixels as the basic processing unit to obtain video saliency results by combining temporal saliency maps and spatial saliency maps. In [6], Chen et al. proposed a video saliency-detection method based on spatial-time fusion and low rank consistency diffusion. Liu et al. [8] proposed an effective spatiotemporal saliency model for unconstrained videos with complicated motion and complex scenes. In [10], Fang et al. proposed a video saliency-detection method based on feature comparison in the compressed domain, which can be more conveniently applied to network-based multimedia fields, such as video redirection, video quality evaluation, etc. Recently, Wang et al. [9] proposed a deep learning based saliency-detection model to detect salient regions in videos via fully convolutional networks. In [15], Le et al. presented an end-to-end 3D fully convolutional network for salient object detection in a video. Zhang et al. [16] proposed a novel attention guided saliency detection network which selectively integrates multi-level contextual information in a progressive manner. Jetley et al. [28] proposed an end-to-end trainable attention model via convolutional neural network (CNN) architectures built for image classification. This method takes 2D feature vectors as

inputs, which forms the intermediate representations of the input image at different stages in the CNN pipeline, and finally outputs a 2D matrix of scores for each saliency map. Jonathan et al. [41] built "fully convolutional" networks that take input of arbitrary sizes to produce correspondingly saliency outputs with efficient inference and learning.

Inspired by the previous works [16, 28], we proposed a novel method for video saliency detection by using full convolutional networks to share attention features. Then the center prior is used to remove the noises in the image backgrounds. Experimental results on two freely available databases verify the designed model is effective.

The rest of the paper is organized as follows. In Section II, we will introduce the proposed framework in detail. Experimental results are presented in Section III. The paper closes with a conclusion and discussion in Section IV.

II. The Proposed Method

A. Deep networks for frame saliency

In this section, we describe the proposed video saliency-detection framework. As depicted in Fig. 2, this video saliency deep-network takes three frames as an input unit and produces three corresponding saliency maps with the same size of the input frame. We model this process with a fully convolutional network (FCN), and there will be a feature weight sharing in the middle. We denote these three frames of the input unit as $F_{t\text{-}1}$, F_t, and F_{t+1} of the FCN network model. Each frame of the video will be extracted its own features through the bottom of the networks, which is a stack of convolutional layers. In order to illustrate the process of being placed in the networks, let's take the F_t frame as input to the FCN as an example, the outputs of each convolutional layer are a set of arrays, called feature maps, with size $h \times w \times c$, where h, w and c are height, width and the feature or channel dimensionality, respectively. In the convolutional layer, a given frame or feature map is used as a new input, and a multi-layer feature map is obtained through a plurality of convolution kernels. Each feature map is obtained by convolving with a trainable linear filter (or kernel) at each position of the input feature map with a trainable bias parameter. If we denote the input feature map as F, whose convolution filters are determined by the kernel weight W_c and bias b_c, then the output feature map is obtained with the following form:

$$y_c = W_c *_s F + b_c, \tag{1}$$

where $*_s$ is the convolution operation, the stride in the convolution process is s.

In the experiments, we have multiple choices on the activation function (e.g., ReLU, Sigmoid) of the convolutional layer. In addition, the form of nonlinear down-sampling (e.g., max pooling) tends to be used after the convolutional layer. Its purpose is to reduce the size of the feature map, simplify network calculation complexity and extract the main visual features of the input frames.

Due to the down-sampling of the convolutional layer and multiple pooling layers, the feature maps generated by the deep network are not sufficient to satisfy the resolution of saliency maps. In other words, the feature map of the output at this time is obscure, and its resolution is lower. To resolve this issue, we perform up-sampling and multi-layer deconvolution during rhe saliency-detection process. After up-sampling, we also weight the corresponding feature map of the convolutional layer before down-sampling. The weight is can be multi-frame shared, which will be described in following section. Thus, we put the up-sampling of the feature map and the multi-layer deconvolution at the top of the deep network model. They can be used the formulas to describe this operation as follows:

$$q_v = Y_c \times V_s, \tag{2}$$

$$y_p = W_d *_s Q_v + b_d, \tag{3}$$

where $\times$ is the multiply operation, the Yc is the last layer that is incorporated into the feature map, V_S is a weight sharing vector between different frames, the stride is set as S. And the kernel weight is W_d and the bias is b_d in the convolution filter, accordingly. In order to ensure the same size between the input and output, we also set the upsampling stride size to S. The prediction-saliency map produced by the current frame through learning is y_p, and Q_v is a feature map obtained after weightsharing processing. The loss function of network model is defined as follows:

$$L(\theta, w) = l_{predict}(\theta, w_{predict}), \tag{4}$$

where θ denotes the collection of all network layer parameters, and w is the weights of the corresponding layer. l is the binary cross-entropy loss, which can be used to balance classes between the predicted saliency $Y \in \{0,1\}^N$ and its corresponding ground truth $G \in \{0,1\}^N$:

$$l = -\sum_{i=1}^{N} \{(1-a) g_i \log y_i + a(1-g_i)\log(1-y_i)\}, \tag{5}$$

where $N = H \times W$ denotes the size of a frame, and g_i belongs to G, y_i belongs to Y, respectively.

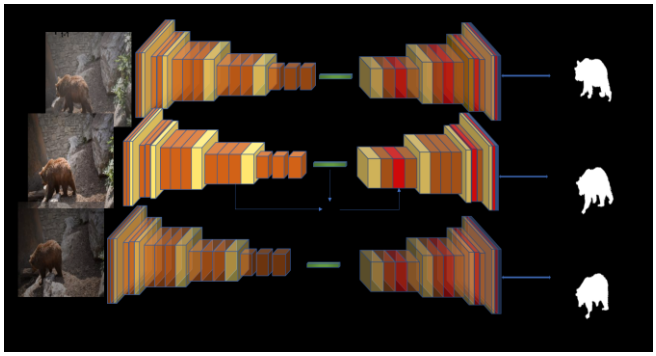


Fig.2. Illustration of our deep network for video saliency detection. Successive three frames ($F_{t\text{-}1}$,F_t ,F_{t+1}) from an input video data are being sent separately into the deep network. The size of the input frame is 224× 224. An interframe vector is generated at the middle of the network to guide saliency-map generation. Finally, a fully convolution network with 1 × 1 kernel and sigmoid activity function are utilzed to output of a probability map with the same size as input, in which larger intensities mean higher saliency values.

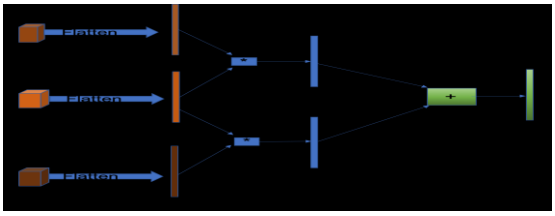


Fig.3. The generated process of the shared vectors, in which the feature map of each frame becomes its corresponding one-dimensional vector by the flatten and dense operation. V_a, V_b, and V_s represent the sharing vectors of two frames and all the three frames, respectively.

B. *Weighting share strategy*

In this section, we will describe the weighting share scheme in the designed video saliency deep network. Because of different frames in a video are not related to each other when they are simultaneously transmitted into the FCN network. However, in practice, the sequence of the video frames is continuous and related. In order to maintain spatial and temporal consistency in the video frames, at the bottom of the network and in the middle of the top network, we add a feature sharing processing, as show in Fig. 3. And the vector in the deep network is thus given by:

$$V_t = D(F_t(y_t);512), \tag{6}$$

$$\begin{aligned} V_a &= V_{t-1} \times V_t, \\ V_b &= V_t \times V_{t+1}, \\ V_s &= V_a + V_b, \end{aligned} \tag{7}$$

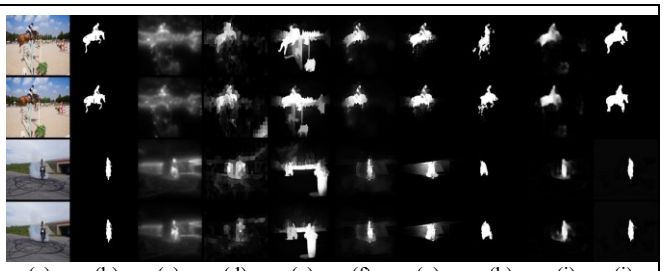
where $F_t(\cdot)$ is flatten treatment and $D(\cdot)$ is densely-connected with 512 dimensionality vector, respectively.

C. *Final Saliency refinement*

At last, the predicted saliency maps can be produced by the deep network model, while the background noises are often generated in the process of learning owing to the backgrounds are also changed due to the salient objects' motion. In order to filter out the background noises and produce a more accurate saliency map, the predicted saliency maps generated by the convolution network are refined with a pixel-by-pixel operation with spatial location cues with center prior of the salient objects [12], as is illustrated in Fig. 1.

III. Experimental Results

For the sake of verifying the performance of the designed saliency-detection model, the publicly available DAVIS datasets [10] were tested for evaluation and comparison. In addition, five typical state-of-the-art saliency-detection methods are chosen, containing saliency-detection methods of random walk with restart (RWRV) [17]; Superpixel-level graph and spatiotemporal propagation (SGSP) [28]; Fully convolutional networks (FCN) [26]; Gradient-flow filtered based model (GAFL) [3]; Fusion and low-rank coherency diffusion (FLRC) [18]; Minimum barrier algorithm (MB) [33]; and 3D fully convolutional network (DSR3D) [15]. In our experiments, for scenarios with large background changes, the deep learning-based approach is more effective.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 4. Comparison of different state-of-the-art saliency detection models based on the DAVIS database. (a) Input images; (b) GT; (c) RWRV; (d) SGSP; (e) MB; (f) GAFL; (g) FLRC; (h) DSR3D; (i) FCN; (j) OUR method.

Fig. 4 illustrates some experimental results of different algorithms performed on the DAVIS database. From Fig. 4 (h), it can be seen that our designed model can produce much more accurate saliency maps with the prominent objects highlighted.

In addition to objectively analysis of different models, we also compared the performance of the proposed method with other five state-of-the-art saliency-detection methods. Fig.5 shows the widely used Mean Absolute Error (MAE) [14] and the *F*-measure [19] values of all the different methods. From comparisons, we can note that our approach is superior to other methods.

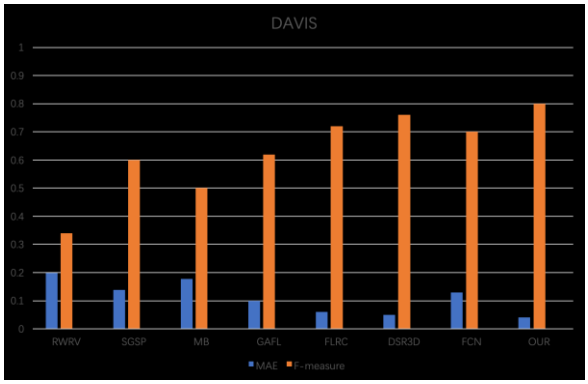


Fig. 5. Comparison of different saliency-detection methods in terms of average MAE and *F*-measure based on the DAVIS dataset.

IV. Conclusion and Discussion

In this paper, we designed a video saliency-detection model based on full convolution neural networks and center prior. With a weighting share strategy, it can capture spatial and temporal features of dynamic scenes. Then, the saliency map estimated from the weighting saliency network is post-treatmented with a center prior, which enables our method to eliminate background noises. The experiments carried on the public video dataset show that our proposed method outperforms the state-of-the-art methods.

Acknowledgment

This work was supported by National Natural Science Foundation of China (NSFC) (61601427, 61602229, 61771230); Royal Society - K. C. Wong International Fellowship (NIF\R1\180909).

REFERENCES

[1] Xi T, Zhao W, Wang H, Lin WS. Salient object detection with spatiotemporal background priors for video. IEEE Trans. on Image Processing, 2017, 26(7): 3425–3436.

[2] Seo HJ, Milanfar P. Static and space-time visual saliency detection by self-resemblance. Journal of Vision, 2009, 9(12): 1–27.

[3] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," IEEE Trans. Image Process., vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[4] Liu Z, Zhang X, Luo SH, Meur OL. Superpixel-Based spatiotemporal saliency detection. IEEE Trans. on Circuits and Systems for Video Technology, 2014, 24(9): 1522–1540.

[5] Le TN, Sugimoto A. Region-Based multiscale spatiotemporal saliency for video. arXiv: 1708. 01589, 2017

[6] Chen CLZ, Li S, Wang YG, Qin H, Hao AM. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Trans. on Image Processing, 2017, 26(7): 3156–3170.

[7] M. Jian, Q. Qi, J. Dong, Y. Yin, K. M. Lam, Integrating QDWD with Pattern Distinctness and Local Contrast for Underwater Saliency Detection, Journal of Visual Communication and Image Representation, Vol. 53, pp. 31–41, 2018.

[8] Liu Z, Li JH, Ye LW, Sun GL, Shen LQ. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. IEEE Trans. on Circuits and Systems for Video Technology, 2017, 27(12): 2527–2542.

[9] Wang WG, Shen JB, Shao L. Video salient object detection via fully convolutional networks. IEEE Trans. on Image Processing, 2018, 27(1): 38–49.

[10] Fang YM, Lin WS, Chen ZZ, Tsai CM, Lin CW. A video saliency detection model in compressed domain. IEEE Trans. on Circuits and Systems for Video Technology, 2014, 24(1): 27–38.

[11] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," IEEE Trans. Image Process., vol. 26, no. 7, pp. 3156–3170, 2017.

[12] M. Jian, W. Zhang, H. Yu, et al. Saliency detection based on directional patches extraction and principal local color contrast. Journal of Visual Communication and Image Representation, 2018, 57: 1-11.

[13] L. Itti, C. Koch and E. Niebur. A model of saliency based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20 (11): 1254-1259.

[14] Federico Perazzi, Philipp Kr¨ahenb¨uhl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In CVPR, pages 733–740, 2012

[15] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In BMVC, 2017..

[16] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang; Progressive Attention Guided Recurrent Network for Salient Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 714-722.

[17] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," IEEE Trans. Image Process., vol. 24, no. 8, pp. 2552–2564, 2015.

[18] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," IEEE Trans. Image Process., vol. 26, no. 7, pp. 3156–3170, 2017.

[19] R. Achanta, S. Hemami, F.o Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In IEEE CVPR, pages 1597–1604, 2009.

[20] E. Rahtu, J. Kannala, M. Salo, et al. Segmenting salient objects from images and videos. In Proc. 11th European Conference on Computer Vision, 2010: 366–379.

[21] Ren ZX, Gao SH, Chia LT, Tsang IWH. Region-Based saliency detection and its application in object recognition. IEEE Trans. on Circuits and Systems for Video Technology, 2014, 24(5): 769–779.

[22] Gu K, Wang SQ, Yang H, Lin WS, Zhai GT, Yang XK, Zhang WJ. Saliency-Guided quality assessment of screen content images. IEEE Trans. on Multimedia, 2016, 18(6): 1098–1110.

[23] Xiao DG, Xin C, Zhang T, Zhu H, Li XL. Saliency texture structure descriptor and its application in pedestrian detection. Ruan Jian Xue Bao/Journal of Software, 2014, 25(3): 675–689(in Chinese with English abstract).

[24] C. Yang, L. Zhang, H. Lu, M. Yang. Saliency Detection via Graph-Based Manifold Ranking. CVPR 2013.

[25] M. Cheng, N. Mitra, X. Huang, et al. Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37 (3): 569–582.

[26] Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE TIP 27(1), 38–49 (2018)

[27] A. Oliva, A. Torralba,M. Castelhano, and J. Henderson. Top down control of visual attention in object detection. In ICIP, volume 1, 2003, 253–256.

[28] Liu, Z., Li, J., Ye, L., Sun, G., Shen, L.: Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. IEEE TCSVT 27(12), 2527–2542 (2017)

[29] Jetley S , Lord N A , Lee N , et al. Learn To Pay Attention[J]. 2018.

[30] Boostrom R . Learning to pay attention 1[J]. International Journal of Qualitative Studies in Education, 1994, 7(1):51-64.

[31] H. Cholakkal, J. Johnson and D. Rajan. A classifier-guided approach for top-down salient object detection. Signal Processing: Image Communication, 2016, 45: 24-40.

[32] M. Jian, K. M. Lam, J. Dong, L. Shen, "Visual-patch-attention-aware Saliency Detection", IEEE Transactions on Cybernetics, Vol. 45, No. 8, pp. 1575-1586, 2015.

[33] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient object detection at 80 fps. In IEEE ICCV, pages 1404–1412, 2015.

[34] R. Achanta, A. Shaji, K. Smith, et al., SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282.

[35] C. Yang, L. Zhang and H. Lu. Graph-regularized saliency detection with convex-hull-based center prior. IEEE Signal Processing Letters, 2013, 20 (7): 637-640.

[36] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in CVPR, 2013.

[37] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In ICCV, 2013, 2976–2983.

[38] W. Zhu, S. Liang, Y. Wei, et al. Saliency optimization from robust background detection. IEEE CVPR, 2014: 2814-2821.

[39] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In ECCV, 2012, vol. 7574, pp. 29–42.

[40] J. Wang, H. Lu, X. Li, et al. Saliency detection via background and foreground seed selection. Neurocomputing, 2015, 152: 359-368.

[41] J. Long, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.